DOCUMENT RESUME

ED 380 473 TM 022 767

AUTHOR Kaiser, Javaid

TITLE The Estimation of Correlation Matrix from Data Having Missing Values.

PUB DATE Aug 94

NOTE 19p.; Paper presented at the Islamic Countries Conference on Statistical Sciences (4th, Lahore, Pakistan, August 27-31, 1994).

PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Comparative Analysis; *Correlation; *Estimation (Mathematics); *Matrices; Monte Carlo Methods; Research Methodology; Sample Size; Statistical Analysis

IDENTIFIERS *Missing Data

ABSTRACT

A Monte Carlo study was conducted to compare the efficiency of Listwise deletion, Pairwise deletion, Allvalue, and Samemean methods in estimating the correlation matrix from data that had randomly occurring missing values. The four methods were compared in a 3x3x4 factorial design representing sample size, proportion of incomplete records in the sample, and the number of missing values per record. Each sample represented an N x 8 data matrix. The Pairwise method was found best in estimating the correlation matrix under all experimental conditions except when the incomplete records had 50% of values missing. In this condition, Listwise deletion was considered a better choice. Allvalue and Samemean methods performed exactly the same way under all experimental conditions, but were found less efficient than the Pairwise method. One table, three figures. (Contains 12 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JAVAID KAISER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Estimation of Correlation Matrix from Data Having Missing Values

By

Javaid Kaiser

ED 380 473

Paper Presented at the Fourth Islamic Countries Conference on Statistical Sciences, held in Lahore, Pakistan. August 27-31, 1994

TM022767

Abstract

A Monte Carlo study was conducted to compare the efficiency of Listwise deletion, Pairwise deletion, Allvalue, and Samemean methods in estimating correlation matrix from data that had randomly occurring missing values. The four methods were compared in a 3x3x4 factorial design representing sample size, proportion of incomplete records in the sample, and the number of missing values per record. Each sample represented N X 8 data matrix. Pairwise method was found as the best method in estimating correlation matrix under all experimental conditions except when incomplete records had 50% missing values. In this condition, listwise deletion was considered a better choice. Allvalue and Samemean methods performed exactly the same way under all experimental conditions but were found less efficient than pairwise method.

The Estimation of Correlation Matrix from Data Having Missing Values

Although presence of missing values in a data matrix is a common occurrence, one should be careful in the selection of statistical procedure that handles them. Usually, a researcher would have three possible choices: (a) discard the records that have missing values, (b) estimate the missing values and, then, treat these estimates as actually observed data, or (c) estimate the correlation matrix from the incomplete data matrix and use this correlation matrix for further analysis. Each alternative has its own merits and demerits. The first option can be exercised only when sample size is large. The second choice uses variable mean or predicted score as estimates of missing values. The third choice is appropriate only when data analysis can be initiated from the correlation matrix instead of raw data. Examples include regression, canonical, discriminant, and factor analysis.

There are four methods that can be used to estimate correlation matrix (Dixon, 1983). The listwise deletion method ignores all the incomplete records and uses the remaining records to compute correlation matrix. The pairwise deletion method uses pairs that have actually observed values on both the variables in computing variances and covariances. The third method computes correlation matrix by extracting maximum information form the incomplete data. This method uses all acceptable values available on a variable to compute variance and all available information on complete pairs to compute covariance. As a result, the degrees of

freedom used in computing variances and covariances may not be the same. The fourth method computes means and variances on all variables using all observed values. Covariances are, then, computed by taking deviations from these means.

The advantage of using the listwise deletion method is that the resulting correlation matrix is always definite positive and can be inverted as required in some complex statistical procedures (Frane, 1977). However, it is not a good choice when large proportion of data has missing values. The pairwise deletion method is considered a better choice when a researcher cannot afford to lose data. Glasser (1964) observed that pairwise method yields consistent estimates of means, variances, and correlation in large samples. Marini, Olsen, and Rubin (1980) and Dixon (1977), however, pointed out that the pairwise method is valid only when the pattern of missing values is random. When the pattern of missing values is systematic, pairwise method adds considerable bias. Based on Monte Carlo study, Hester (1976) also found pairwise method a better choice but noted that it does not often lead to a positive definite covariance matrix particularly in small samples. Griliches, Hall, and Hausman (1978) used pairwise and listwise deletion methods in the regression analysis. They found that in spite of using more information, pairwise deletion method lead to the same interpretation of regression results as listwise deletion method.

In terms of the discrepancy between true and estimated values of correlation, Wolf, Bearman, and Flesher (1979) found the ranking

of these four methods in the reverse order. However, they discovered that the stability of ranking depends on the size of true correlations and the proportion of missing values. The standard error of the discrepancy was the lowest for pairwise method and highest for the deletion method. Glasser (1964), and Haitovsky (1968) concluded that the closeness of obtained estimates to the actual ones, had there been no missing values, depends on the size of intercorrelation among variables and the pattern of missing values. Other factors that may have a bearing are the sample size, the proportion of incomplete records in a sample, and the proportion of missing values in a record.

The problem associated with the remaining two methods is their liberal use of all the acceptable values in computing variances and covariances. They are not restricted to using the same records and therefore, the resulting correlation coefficients may exceed unity (Hill, 1981; Frane, 1977). Moreover, negative eigenvalues arise when incomplete records are retained in the data matrix and prevent the matrix from being inverted. In order to make sure that the computed correlation matrix is always semi-definite positive, the matrix with negative eigenvalues should be recomputed by using only positive eigenvalues and their vectors.

The present study was designed to investigate the third option of data analysis. The relative efficiency of various methods in reproducing the correlation matrix of the complete sample from the incomplete data has been determined. Although properties of some of these methods have been explored analytically, the present study

provides empirically evidence of these properties under various experimental conditions.

Method

A Monte Carlo study was conducted using 3x3x4 factorial design. The factors studied were (a) sample size, (b) proportion of incomplete records in a sample (PIR), and (c) number of missing values on an incomplete record (NMV). The methods studied were (a) listwise deletion, (b) pairwise deletion, (c) using all observed values on a variable to compute variance and all available information on pairs to compute covariance (ALLVALUE), and (d) computing means and variances on variables from all the observed values and computing covariances by taking deviations from these means (SAMEMEAN). Methods were compared in terms of the discrepancy between true and estimated correlation matrix.

An 8x8 correlation matrix was selected from literature (See Table 1) to represent multivariate normal population for this study. The random samples of multivariate normal deviates were generated from this population according to the cell specifications of the design matrix. Missing values were artificially created by using random number generator. At first, p proportion of records were selected randomly and then for each selected record, m missing values were created at random. The resulting data matrix with randomly occurring missing values was used by all the four methods to compute correlation matrix.

The statistic D, originally proposed by Timm (1970) and later modified by Gleason and Staelin (1975), was used to determine the relative efficiency of the four methods in retaining the complete sample covariance structure. The statistic D is the root-mean-square deviation of the off-diagonal elements of the two correlation matrices. This procedure was repeated 500 times for each cell of the design matrix. The means and variances of observed discrepancies determined the relative efficiency of the methods in retaining the complete sample covariance structure.

Results

The results obtained on samples of size 30 revealed that listwise deletion method creates the largest discrepancy between the estimated and actual correlation matrix when 10% of records have missing values. Pairwise method produced the smallest D value. Allvalue and Samemean methods produced similar results. However, they were slightly worse in performance than pairwise method but more efficient than listwise deletion method (See Figure 1). The trend lines further indicated that the efficiency of methods in estimating correlation matrix is hampered when incomplete records have more than one missing value.

When 20% of the records had missing values, the relative ranking of methods stayed the same but the efficiency of estimating correlation matrix closer to the original one decreased for all methods. The value of D increased from .055 to .084 for deletion

method and from .025 to .039 for the pairwise method. Listwise deletion method gave the worst performance when 30% of the records were incomplete. The value of D, in this instance, increased to .11. In terms of efficiency, all methods retained their ranks across all three levels of incomplete records and across all levels of missing values. Pairwise, Allvalue, and Samemean methods deteriorated in their efficiency as the number of missing values on incomplete records approached 50%. At this point, no substantial differences were observed between deletion method and the remaining three methods.

Figure 2 displays the efficiency of methods in estimating correlation matrix in samples of size 60. Again, pairwise method produced correlation matrix closest to the original one while deletion method gave the worst performance. This was true at all three levels of incomplete records and at all four levels of missing values. The D values for all methods, at each level of the proportion of incomplete records and the number of missing values, decreased compared to the values observed respectively in samples of size 30. The D values, however, increased whenever the proportion of incomplete records or the number of missing values on incomplete records increased. Moreover, the increase in number of missing values deteriorated the efficiency of methods to the extent that no substantial difference was left among the four methods when 50% of the values on incomplete records were missing.

When correlation matrix was computed using samples of size 120, all methods appeared more efficient compared to their

performance with samples of size 30 and 60. It is evident by D values that were lowest compared to the respective values in smaller samples. The data revealed that pairwise deletion method is still the best method to use when 10%, 20%, or 30% of the sample has incomplete records and the number of missing values on each incomplete record does not exceed 50%. When half of the values on incomplete records are unobserved, all methods performed similarly regardless of the proportion of incomplete records in the sample. The D values, as in previous cases, increased with the increase in the proportion of incomplete records in the sample or when the number of missing values increased on the incomplete records.

Discussion

The results of this Monte Carlo study clearly identify pairwise deletion method as the best method to estimate correlation matrix from data with missing values provided the missing values on each incomplete record do not exceed 50%. This finding was found true in samples of size 30, 60, and 120 and at all three levels of the proportion of incomplete records representing 10%, 20%, and 30% incomplete records. These results are consistent with earlier findings of Glasser (1964), Hester (1976), and Marini et al., (1980) that pairwise method yields better estimates of correlation matrix but contradicted the findings of Raymond (1986) who found pairwise method least effective. The results of this study also corroborate with the findings of Griliches et al. (1978) and

Raymond (1986) who did not find any difference between listwise and pairwise deletion methods. However, in the present study, this was true only when 50% of the values on incomplete records were missing. The data also revealed that the efficiency of methods in computing correlation matrix closer to the original one decreases with the increase in the number of missing values or when the proportion of incomplete records increase. Although listwise deletion method has been found inferior to pairwise, Allvalue and Samemean methods, no substantial differences existed among these methods when 50% of the values were missing on each incomplete record. In such situation, listwise deletion method seemed a better choice as marginal efficiency of pairwise method can be sacrificed for computational ease of deletion method. Although all four methods become more efficient when sample size increased, the relative ranking of these methods stayed the same al all levels of sample size, the proportion of incomplete records, and the number of missing values.

Conclusion

The results of this study revealed that pairwise method is the best method to estimate correlation matrix when data has randomly occurring missing values. The efficiency of pairwise method increased with the increase in sample size but decreased with the increase in the number of missing values or the proportion of incomplete records. The deletion method was considered a better choice when 50% values on the incomplete records are missing.

References

Dixon, W. J. (1977), (Eds.). BMDP Biomedical Computer Programs: P-Series. Barkley: University of California Press.

Frane, J. W. (1977). Description and estimation of missing data. In W. J. Dixon and M. B. Brown (1977), (Eds.). BMDP Biomedical Computer Programs: P-Series. Barkley: University of California Press.

Glasser, M. (1964). Linear regression analysis with missing observation among the independent variables. Journal of the American Statistical Association, 59, 834-844.

Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. Psychometrika, 40 (2), 229-252.

Griliches, Z., Hall, B. H., & Hausman, J. A. (1978). Missing data and self-selection in large panels. Annals de L'insee, 30-31, 137-176.

Haitovsky, Y. (1968). Missing data in regression analysis. Journal of the Royal Statistical Society, Series B, 30, 67-82.

Hester, D. B. (1976). Estimation from Incomplete Samples. Madison, Wisconsin: Social Systems Research Institute, University of Wisconsin.

Hill, MaryAnn (1981). Annonated Computer Output for Data Screening. Technical Report No. 77. Veteran's Administration Medical Center, Brentwood and BMDP Statistical Software, Department of Biomathematics, UCLA.

Marini, M. M., Olsen, A. R., & Rubin, D. B. (1980). Maximum likelihood estimation in panel studies with missing data. In Karl F. Schussler, Sociological Methodology-1980. San Francisco: Jossey-Boss Publications.

Raymond, M. R. (1986). Missing data in evaluation research. Evaluation and the Health Professions, 9 (4), 395-420.

Timm, H. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika, 35 (4), 417-438.

Wolfe, B. L., Behrman, J. R., & Flesher, J. (1979). A Monte Carlo Study of Alternative Approaches for Dealing with Randomly Missing Data. Report No. DP# 587-79. Institute for Research on Poverty, University of Wisconsin-Madison.

Table 1

The statistic D for methods at various levels of sample size (N), proportion of incomplete recods (PIR), and the number of missing values per record (NMV)

N	PIR	NMV	Pairwise	Allvalue	Listwise	Samemean
30	0.1	1	0.0251	0.0270	0.0546	0.0270
		2	0.0353	0.0379	0.0542	0.0379
		3	0.0425	0.0449	0.0538	0.0449
		4	0.0483	0.0509	0.0551	0.0509
	0.2	1	0.0385	0.0413	0.0839	0.0413
		2	0.0538	0.0575	0.0841	0.0575
		3	0.0657	0.0700	0.0849	0.0700
		4	0.0731	0.0770	0.0840	0.0770
	0.3	1	0.0491	0.0527	0.1136	0.0526
		2	0.0682	0.0728	0.1096	0.0727
		3	0.0838	0.0893	0.1106	0.0893
		4	0.0949	0.1008	0.1113	0.1007
60	0.1	1	0.0178	0.0193	0.0379	0.0193
		2	0.0251	0.0270	0.0376	0.0270
		3	0.0298	0.0318	0.0375	0.0318
		4	0.0328	0.0349	0.0377	0.0349
	0.2	1	0.0266	0.0287	0.0574	0.0287
		2	0.0372	0.0401	0.0578	0.0401
		3	0.0437	0.0469	0.0569	0.0469
		4	0.0503	0.0535	0.0577	0.0535
	0.3	1	0.0332	0.0361	0.0774	0.0361
		2	0.0467	0.0505	0.0756	0.0505
		3	0.0575	0.0617	0.0769	0.0617
		4	0.0657	0.0701	0.0769	0.0701
120	0.1	1	0.0125	0.0136	0.0266	0.0136
		2	0.0174	0.0189	0.0264	0.0189
		3	0.0211	0.0226	0.0270	0.0226
		4	0.0235	0.0250	0.0266	0.0250
	0.2	1	0.0186	0.0202	0.0404	0.0202
		2	0.0260	0.0282	0.0401	0.0282
		3	0.0313	0.0338	0.0403	0.0338
		4	0.0350	0.0375	0.0402	0.0375
	0.3	1	0.0231	0.0252	0.0526	0.0252
		2	0.0326	0.0355	0.0527	0.0355
		3	0.0398	0.0432	0.0533	0.0432
		4	0.0454	0.0488	0.0536	0.0488

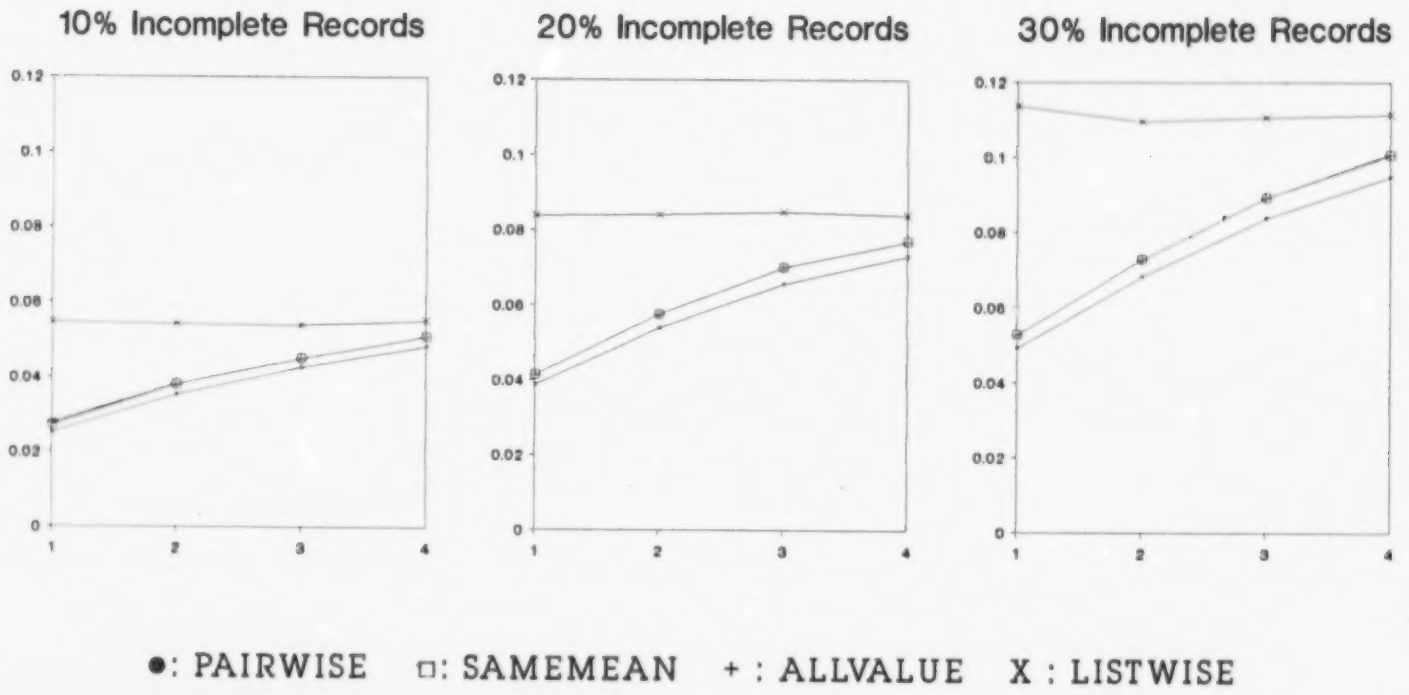

Figure 1. Statistic D for samples of size 30.

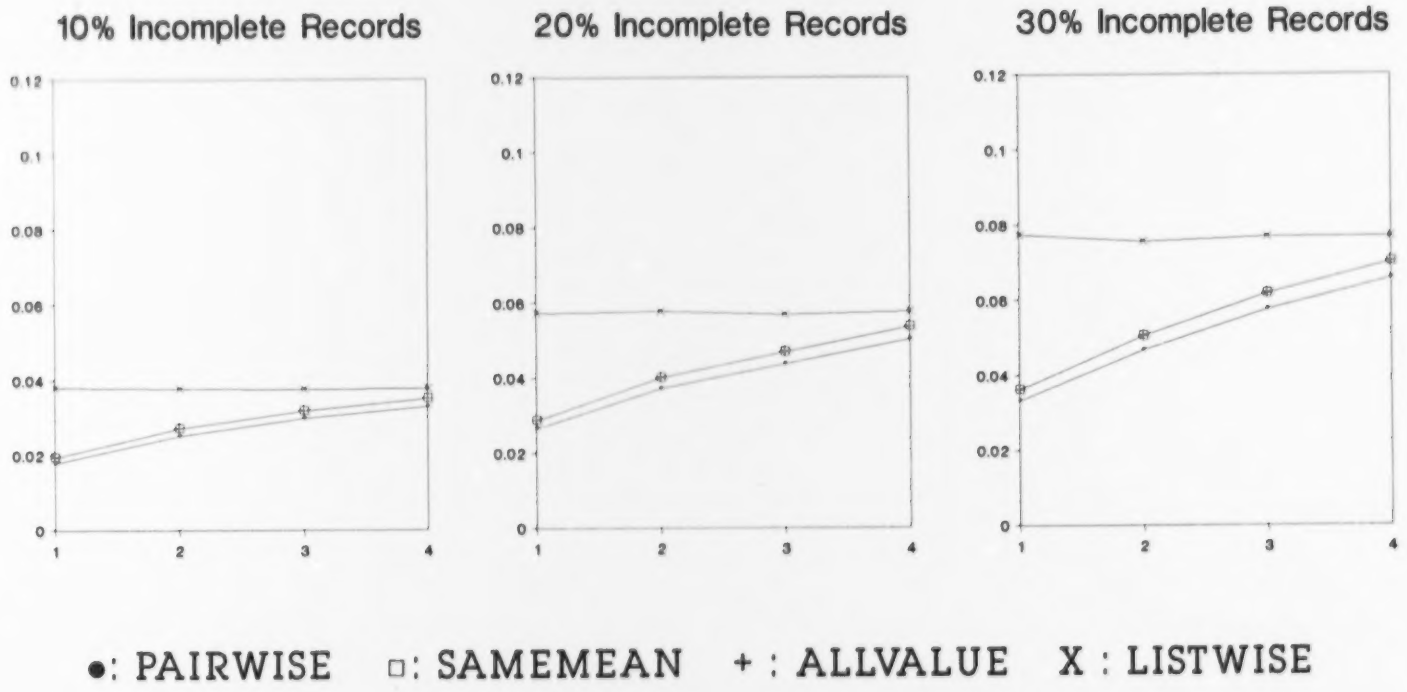

Figure 2. Statistic D for samples of size 60.

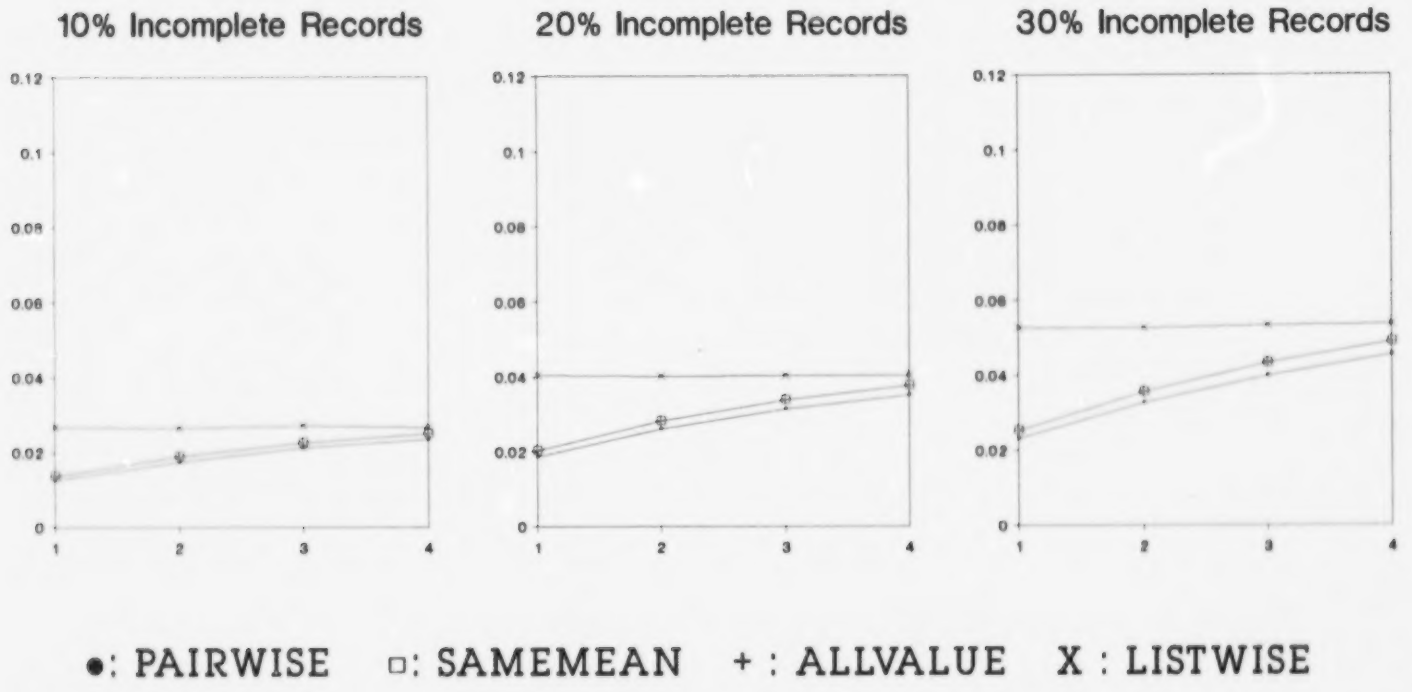

Figure 3. Statistic D for samples of size 120.